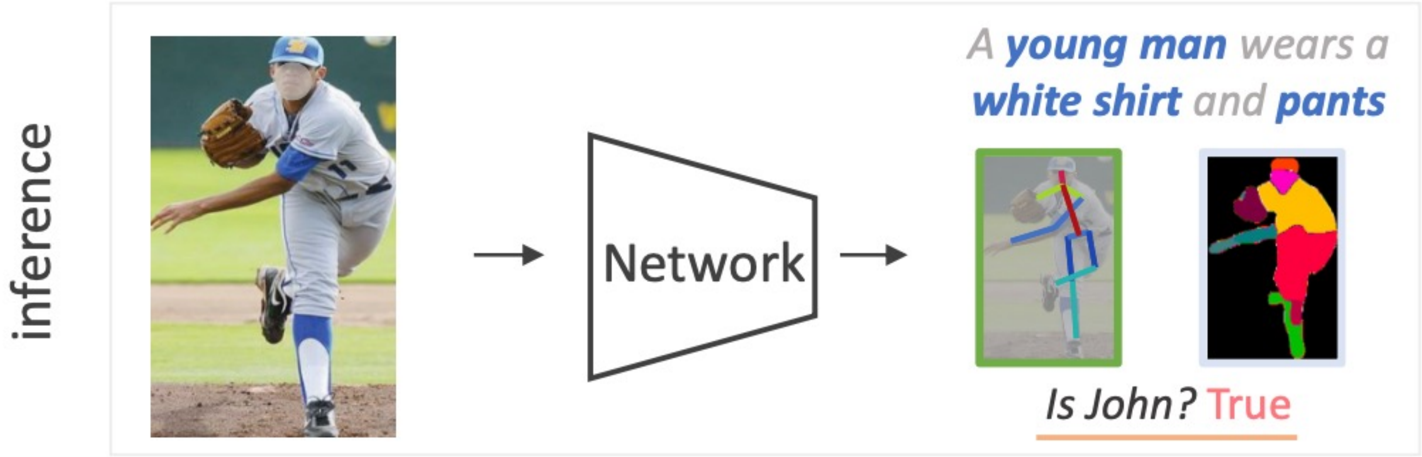
Split to Learn: Gradient Split for Multi-Task Human Image Analysis

Weijian Deng[1] Yumin Suh[2] Xiang Yu[2] Masoud Faraki[2]
Liang Zheng[1] Manmohan Chandraker[2,3]

[1]Australian National University [2]NEC Labs America
[3]University of California, San Diego

Multi-Task Human Image Analysis

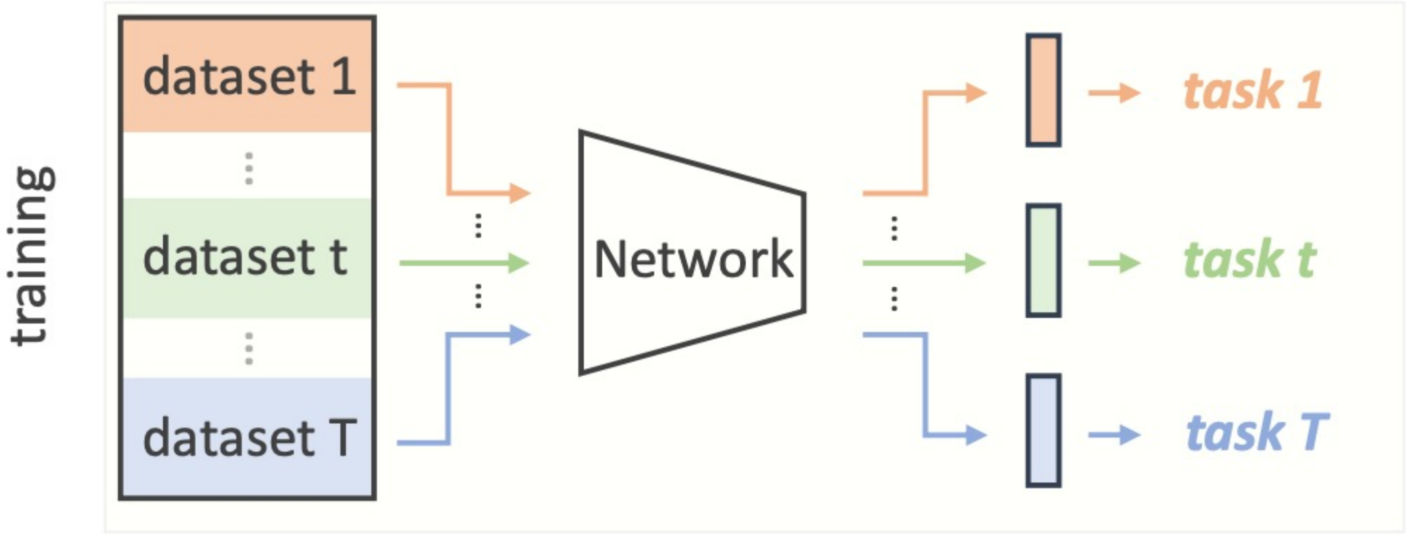
Multi-task network provides a **rich explanation** of person-body images, including attributes, pose, part masks, and identity



Multi-Task Human Image Analysis

Practical setting

Multi-task networks are trained across datasets and each dataset does not necessarily have exhaustive annotations for all tasks



Task Conflict

Multi-task learning can encounter **task conflicts**

- ✓ Identity-variance vs. identity-invariance
  - Attribute recognition vs. Pose estimation
- ✓ Pose-variance vs. Pose invariance
  - Pose estimation vs. Person re-identification
- ✓ …

Task Conflict

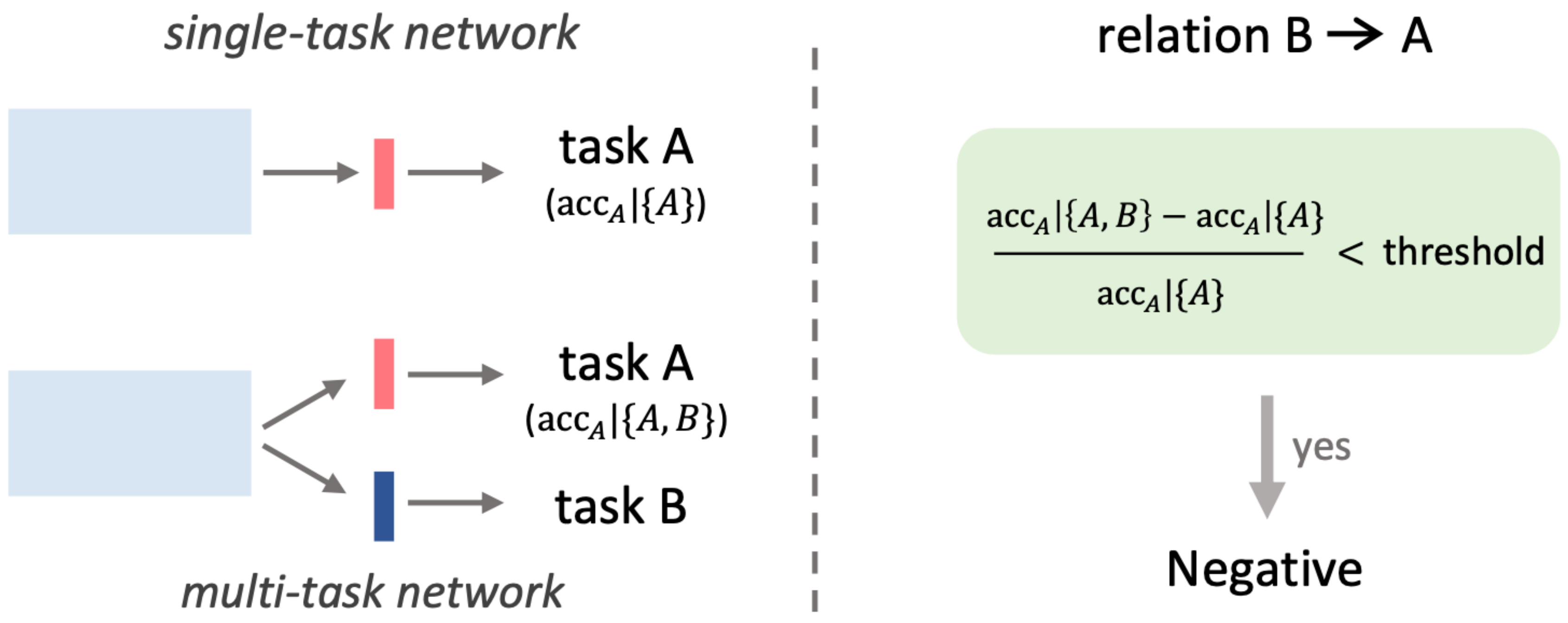
Multi-task learning can encounter **task conflicts**

- **Our goal** is to train a unified model that solves multiple human-related tasks while avoiding the task conflict

better accuracy-efficiency trade-off

Gradient Split

Asymmetric Inter-task relation definition



relative accuracy change

Gradient Split

Asymmetric Inter-task relation definition

Trained With	Relative Performance Change On: Attribute	ReID	Pose	Parsing
Attribute	—	-2.16%	-1.47%	-9.87%
ReID	-2.05%	—	-1.36%	-16.22%
Pose	-0.77%	-0.86%	—	0.00%
Parsing	-0.91%	-0.97%	0.11%	—

Threshold: -0.01

Gradient Split

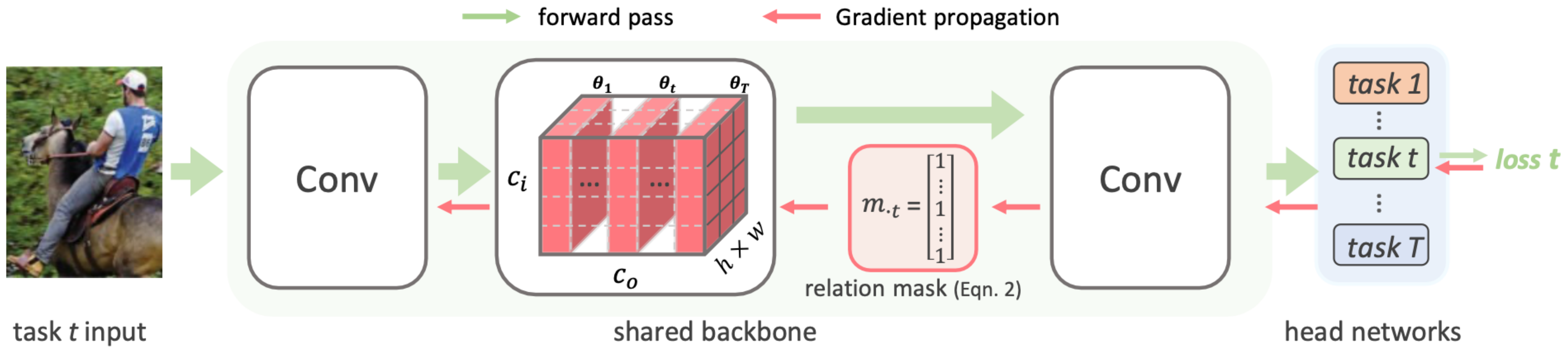
Asymmetric Inter-task relation definition

		Performance On			
		Attribute	ReID	Pose	Parsing
Trained With	Attribute	—	↓	↓	↓
	ReID	↓	—	↓	↓
	Pose	—	—	—	—
	Parsing	—	—	—	—

↓ Negative relation

Gradient Split

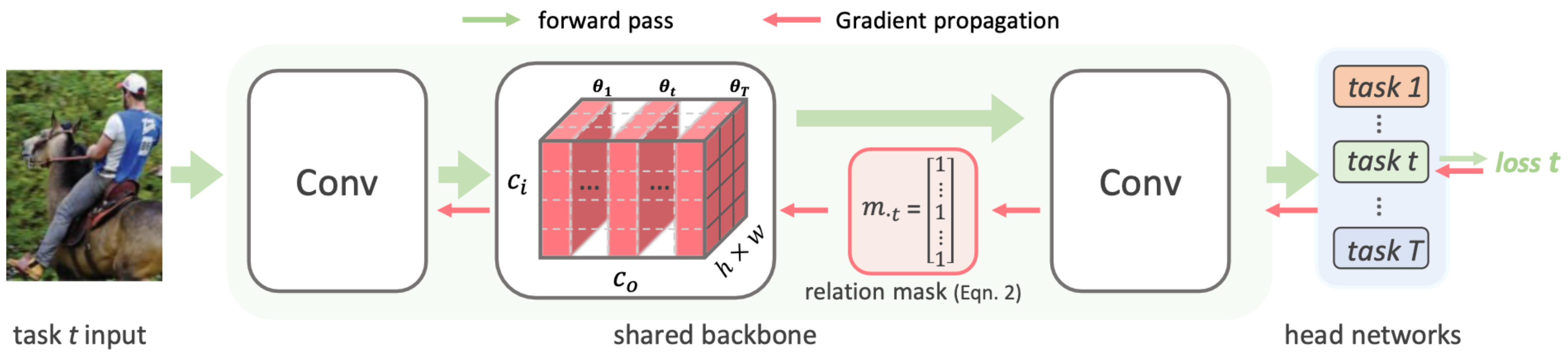
Framework



Multi-head framework for multi-task learning

Gradient Split

Framework

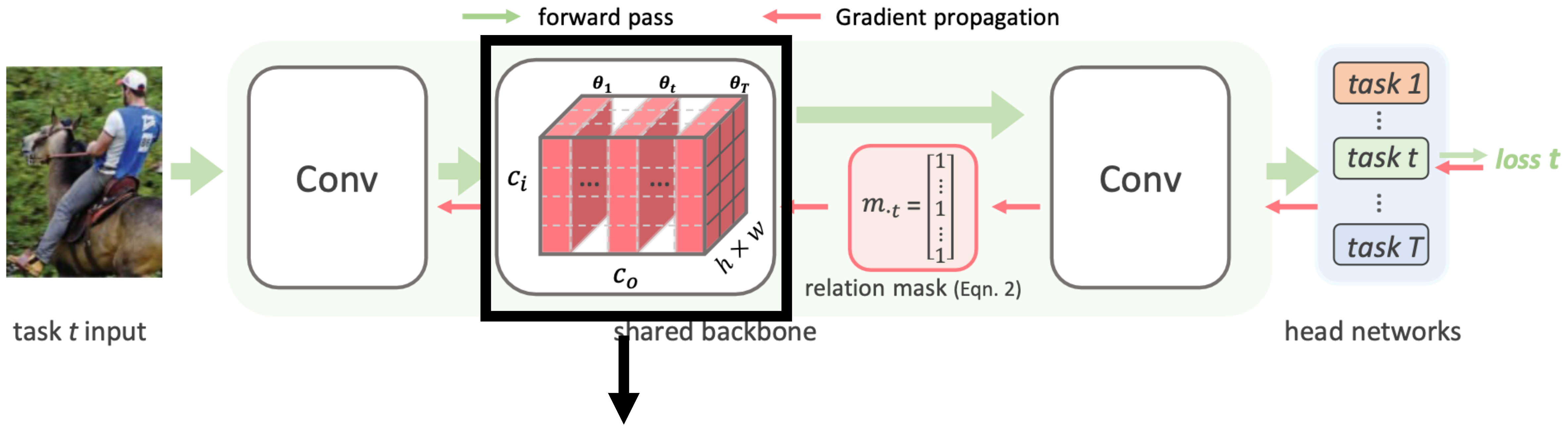


Gradient split is only conducted during the backward process

No extra forward cost and No network change

Gradient Split

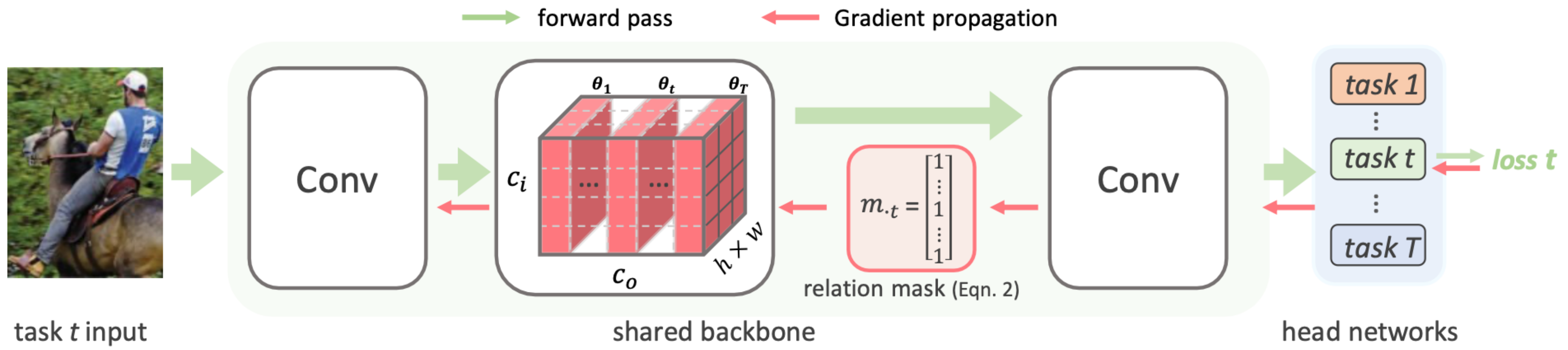
Inter-task Relationship based Gradient Update



We divide parameters of shared backbone into T groups for T tasks

Gradient Split

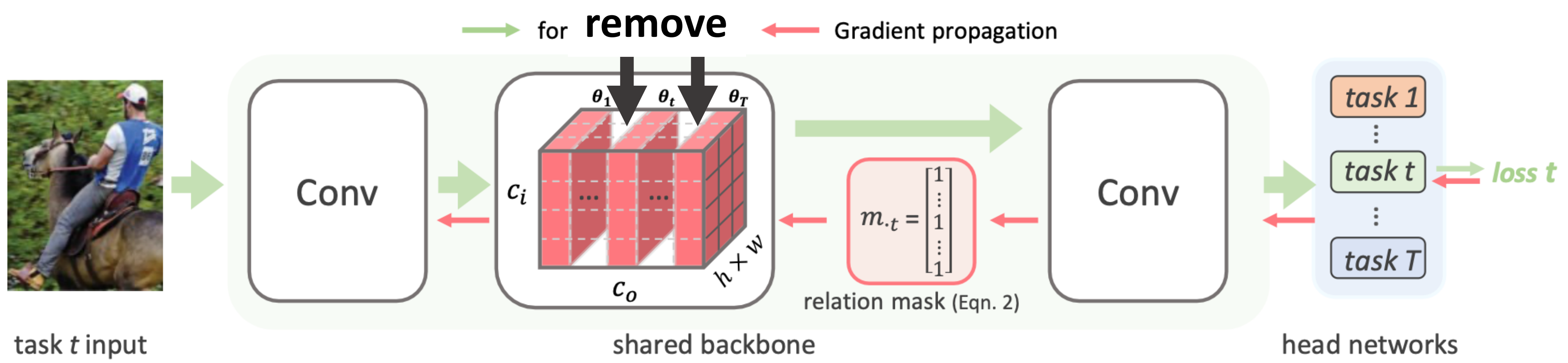
Inter-task Relationship based Gradient Update



GradSplit updates parameter θ_t using the gradients from only a subset of tasks $\{t'\}$, where the relationship task $t' \rightarrow t$ is not negative, while discarding gradients from the other tasks.

Gradient Split

Inter-task Relationship based Gradient Update



GradSplit updates parameter θ_t using the gradients from only a subset of tasks $\{t'\}$, where the relationship task $t' \rightarrow t$ is not negative, while discarding gradients from the other tasks.

Experiment: Four-Task Analysis

Methods	Backbone	ReID mAP (↑)	Attribute MA (↑)	Pose Mean (↑)	Parsing mIoU (↑)	Δ_m (↑)	#Param (M) ↓	#FLOPs (G) ↓
Single-task Networks (Upperbound)	ResNet-50-GN	81.1	78.0	88.2	45.6	+0.0	123	41
	ResNet-50-BN	83.0	78.3	88.4	45.4	–	123	41
Single-task Networks (Baseline)	ResNet-18-GN	74.9	76.9	87.0	42.4	–	63	24
	ResNet-18-BN	74.2	74.2	87.4	41.9	–	63	24
RCM [15]	ResNet-50-GN	54.9	68.1	69.0	36.1	-21.9	141	80
SFG [2]	ResNet-50-GN	64.4	73.9	71.8	34.8	-17.0	52	20
GradNorm [4]	ResNet-50-GN	56.1	77.7	68.4	28.5	-23.1	52	18
MTAN [21]	ResNet-50-GN	42.7	77.4	86.0	41.9	-14.7	75	40
ASTMT [26]	ResNet-50-TBN*	50.6	78.9	87.0	43.6	-10.6	82	42
Multi-head Baseline	ResNet-50-BN	63.2	76.3	78.9	39.8	-11.9	52	18
	ResNet-50-TBN*	78.1	77.2	86.8	41.8	-3.7	52	41
	ResNet-50-GN	79.3	76.4	86.1	42.7	-3.3	52	18
GradSplit (Ours)	ResNet-50-GN	80.1	77.8	86.4	43.9	**-1.8**	52	18

Experiment: Four-Task Analysis

Methods	Backbone	ReID mAP (↑)	Attribute MA (↑)	Pose Mean (↑)	Parsing mIoU (↑)	Δ_m (↑)	#Param (M) ↓	#FLOPs (G) ↓
Single-task Networks (Upperbound)	ResNet-50-GN	81.1	78.0	88.2	45.6	+0.0	123	41
	ResNet-50-BN	83.0	78.3	88.4	45.4	–	123	41
Single-task Networks (Baseline)	ResNet-18-GN	74.9	76.9	87.0	42.4	–	63	24
	ResNet-18-BN	74.2	74.2	87.4	41.9	–	63	24
GradNorm [4]		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
MTAN [21]		42.7	77.4	86.0	41.9	-14.7	75	40
ASTMT [26]	ResNet-50-TBN*	50.6	78.9	87.0	43.6	-10.6	82	42
Multi-head Baseline	ResNet-50-BN	63.2	76.3	78.9	39.8	-11.9	52	18
	ResNet-50-TBN*	78.1	77.2	86.8	41.8	-3.7	52	41
	ResNet-50-GN	79.3	76.4	86.1	42.7	-3.3	52	18
GradSplit (Ours)	ResNet-50-GN	80.1	77.8	86.4	43.9	**-1.8**	52	18

GradSplit achieves a better **accuracy-efficiency trade-off**

Experiment: Three-Task Analysis

Pose + Attribute + ReID

Methods	Backbone	Attribute MA (↑)	ReID mAP (↑)	Pose Mean (↑)	Δ_m (↑)	#Param (M) ↓
Single-task	R50-GN	78.0	81.1	88.2	+0.0	85
	R18-GN	76.9	74.9	87.0	–	39
Cross-stitch [27]	R18-GN	76.3	72.7	86.8	-4.7	38
NDDR [10]	R18-GN	76.1	69.3	86.8	-6.2	42
GradNorm [4]	R50-GN	74.0	54.5	85.1	-13.8	38
MTAN [21]	R50-GN	77.4	50.0	85.5	-14.0	38
Multi-head	R50-GN	75.9	76.5	86.3	-3.5	38
GradSplit	R50-GN	77.6	80.2	86.3	**-1.3**	38

GradSplit is more effective than other methods

Experiment: Large Capacity Backbone

Methods	Backbone	Attr	ReID	Pose	Parsing	Δ_m	#Param
		MA	mAP	Mean	mIoU	(↑)	(M)
Single-task	R50-GN	78.0	81.1	88.2	45.6	+0.0	123
Task-specific L4	R50-L4	76.8	78.2	86.4	43.5	-2.9	96
DropGrad (p=0.50)	R50-GN+	77.9	80.2	86.4	42.2	-2.7	72
Multi-head		77.1	80.4	87.8	46.9	+0.1	72
GradSplit		78.2	81.6	87.9	47.4	**+1.1**	72

GradSplit outperforms the Single-task networks

GradSplit achieves the best accuracy-efficiency trade-off

Experiment: Which Layer?

Methods		Pose	Attribute	ReID		Parsing
		Mean	MA	Rank-1	mAP	mIoU
Multi-head Basel.		84.9	75.5	86.2	64.7	38.0
GradSplit	Layer 4	**85.4**	**77.1**	**89.2**	**71.4**	**39.1**
	Layer 3-4	85.0	77.1	88.0	68.0	38.3
	Layer 2-4	85.2	77.0	87.4	67.6	38.0
	Layer 1-4	84.6	77.0	87.6	66.9	36.6

Last Layer is best choice

Different tasks might share the common features in previous layers

Experiment: Random Drop?

Methods		Pose	Attribute	ReID		Parsing
		Mean	MA	Rank-1	mAP	mIoU
Multi-head Basel.		84.9	75.5	86.2	64.7	38.0
GradSplit	Layer 4	**85.4**	**77.1**	**89.2**	**71.4**	**39.1**
	Layer 3-4	85.0	77.1	88.0	68.0	38.3
	Layer 2-4	85.2	77.0	87.4	67.6	38.0
	Layer 1-4	84.6	77.0	87.6	66.9	36.6
DropGrad (p=0.50)		81.5	74.0	85.8	64.3	36.3
DropGrad (p=0.75)		81.5	73.9	85.3	63.7	36.8

Randomly drop gradients **does not** help

Thank you